

Characterising variation of nonparametric random probability measures using the Kullback-Leibler divergence

JAMES WATSON¹, LUIS NIETO-BARAJAS^{1,2} and CHRIS HOLMES¹

¹*Department of Statistics, University of Oxford, UK*

²*Department of Statistics, ITAM, Mexico*

Abstract

This work studies the variation in Kullback-Leibler divergence between random draws from some popular nonparametric processes and their baseline measure. In particular we focus on the Dirichlet process, the Pólya tree and the frequentist and Bayesian bootstrap. The results shed light on the support of these nonparametric processes. Of particular note are results for finite Pólya trees that are used to model continuous random probability measures. Our results provide guidance for specifying the parameterisation of the Pólya tree process that allows for greater understanding while highlighting limitations of the standard canonical choice of parameter settings.

Keywords: Bayesian nonparametrics, Kullback-Leibler divergence, bootstrap methods, Pólya trees.

1 Introduction

Random probability models are key components of Bayesian nonparametrics (Hjort *et al.*, 2010; Ghosh & Ramamoorthi, 2003; Müller & Quintana, 2004) used to express prior beliefs with wide support. Bayesian nonparametrics has become increasingly popular in recent years due to the flexible modelling

structures it supports and alleviating concerns over the “closed hypothesis space” of Bayesian inference. The most commonly used processes are the Dirichlet process prior or generalizations of it, and the Pólya tree prior (PT), which includes the Dirichlet process as a special case, some of the main references being: Ferguson (1973); Lavine (1992); Hjort *et al.* (2010). Both the Dirichlet and the Pólya tree priors are of particular interest because of their analytical tractability and their conjugacy properties for inference problems.

The properties of these processes are usually given at the level of characterizing their mean and variance when defining the process around a particular centring distribution F_0 . For instance, if we have a random distribution F with a Dirichlet process law, denoted $F \sim \mathcal{DP}(\alpha, F_0)$, where F_0 is the centring distribution, then $E(F) = F_0$, and α is a precision parameter that controls the dispersion of F from F_0 . Similarly, if F is a random distribution with law governed by a Pólya tree process, using notation from Hanson (2006), $F \sim \mathcal{PT}(\alpha, \rho, F_0)$, where ρ denotes the precision function, then selecting α and a partition structure Π that defines the tree, the draws will be centred around F_0 , and α is again the precision parameter. Moreover, the precision function $\rho(\cdot)$ controls the speed at which the variance of the branching probabilities that define the PT increase or decrease. Lavine (1992) recommends $\rho(m) = m^2$ as a “sensible canonical choice”, which has been adopted as the standard choice in the vast majority of applications, see for example Karabatsos (2006); Muliere & Walker (1997); Walker *et al.* (1999); Hanson & Johnson (2002); Walker & Mallick (1997). In practical applications when using Pólya trees in Bayesian inference for example, it is also necessary to truncate the tree at a certain level M . One consequence of our work allows better insight for both choosing the truncation level M and for choosing the function parameter $\rho(\cdot)$.

More generally we consider the general question of how far a random draw F , is from a specific centring distribution F_0 . We also ask ourselves whether it is possible to set the parameters of the model in order to sample distributions at a specific divergence from F_0 . In this note we provide some guidance on how to answer these questions using the most common measure of divergence between densities, the Kullback-Leibler (KL) divergence (Kullback & Leibler, 1951). We concentrate on this divergence for its fundamental role played in information theory and Bayesian statistics (e.g. Kullback, 1997; Bernardo & Smith, 1994; Cover & Thomas, 1991).

Section 2 introduces some notation and defines the Pólya tree as the

principal model considered. We also consider the Bayesian and frequentist bootstrap procedures in section 3. Section 3 presents several properties of the KL divergence, considering random draws of some random probability models. Section 4 concludes with a discussion on the implications of these findings.

2 Notation

The Pólya tree will be our main object of interest, particularly as the Dirichlet process can be seen as a particular case of a Pólya tree, see Ferguson (1974). We define it as follows.

The Pólya tree relies on a binary partition tree of the sample space. For simplicity of exposition we consider $(\mathbb{R}, \mathcal{B})$ as our measurable space with \mathbb{R} the real line and \mathcal{B} the Borel sigma algebra of subsets of \mathbb{R} . Using the notation in Nieto-Barajas & Müller (2012), the binary partition tree is denoted by $\Pi = \{B_{mj} : m \in \mathbb{N}, j = 1, \dots, 2^m\}$, where the index m denotes the level in the tree and j the location of the partitioning subset within the level. The sets at level 1 are denoted by (B_{11}, B_{12}) ; the partitioning subsets of B_{11} are (B_{21}, B_{22}) , and $B_{12} = B_{23} \cup B_{24}$, such that $(B_{21}, B_{22}, B_{23}, B_{24})$ denote the sets at level 2. In general, at level m , the set B_{mj} splits into two disjoint sets $(B_{m+1,2j-1}, B_{m+1,2j})$, where $B_{m+1,2j-1} \cap B_{m+1,2j} = \emptyset$ and $B_{m+1,2j-1} \cup B_{m+1,2j} = B_{mj}$.

We associate random branching probabilities Y_{mj} with every set B_{mj} . We will use F to denote a cdf or a probability measure in-distinctively, and f to denote a density. We define $Y_{m+1,2j-1} = F(B_{m+1,2j-1} \mid B_{mj})$, and $Y_{m+1,2j} = 1 - Y_{m+1,2j-1} = F(B_{m+1,2j} \mid B_{mj})$. We denote by $\mathcal{Y} = \{Y_{mj}\}$ the set of random branching probabilities associated with the elements of Π .

DEFINITION 1 (Lavine, 1992). Let $\mathcal{A}_m = \{\alpha_{mj}, j = 1, \dots, 2^m\}$ be non-negative real numbers, $m = 1, 2, \dots$, and let $\mathcal{A} = \bigcup \mathcal{A}_m$. A random probability measure F on $(\mathbb{R}, \mathcal{B})$ is said to have a Pólya tree prior with parameters (Π, \mathcal{A}) , if for $m = 1, 2, \dots$ there exist random variables $\mathcal{Y}_m = \{Y_{m,2j-1}\}$ for $j = 1, \dots, 2^{m-1}$, such that the following hold:

- (i) All the random variables in $\mathcal{Y} = \bigcup_m \mathcal{Y}_m$ are independent.
- (ii) For every $m = 1, 2, \dots$ and every $j = 1, \dots, 2^{m-1}$, $Y_{m,2j-1} \sim \text{Be}(\alpha_{m,2j-1}, \alpha_{m,2j})$.

(iii) For every $m = 1, 2, \dots$ and every $j = 1, \dots, 2^m$

$$F(B_{mj}) = \prod_{k=1}^m Y_{m-k+1, j_{m-k+1}^{(m,j)}},$$

where $j_{k-1}^{(m,j)} = \lceil j_k^{(m,j)} / 2 \rceil$ is a recursive decreasing formula, whose initial value is $j_m^{(m,j)} = j$, that locates the set B_{mj} with its ancestors upwards in the tree. $\lceil \cdot \rceil$ denotes the ceiling function, and $Y_{m,2j} = 1 - Y_{m,2j-1}$ for $j = 1, \dots, 2^{m-1}$.

There are several ways of centring the process around a parametric probability measure F_0 . The simplest and most used method (Hanson & Johnson, 2002) consists of matching the partition with the dyadic quantiles of the desired centring measure and keeping α_{mj} constant within each level m . More explicitly, at each level m we take

$$B_{mj} = \left(F_0^{-1} \left(\frac{j-1}{2^m} \right), F_0^{-1} \left(\frac{j}{2^m} \right) \right], \quad (1)$$

for $j = 1, \dots, 2^m$, with $F_0^{-1}(0) = -\infty$ and $F_0^{-1}(1) = \infty$. If we further take $\alpha_{mj} = \alpha_m$ for $j = 1, \dots, 2^m$ we get $E\{F(B_{mj})\} = F_0(B_{mj})$.

In particular, we take $\alpha_{mj} = \alpha \rho(m)$, so that the parameter α can be interpreted as a precision parameter of the Pólya tree (Walker & Mallick, 1997), and the function ρ controls the speed at which the variance of the branching probabilities moves down in the tree. According to Ferguson (1974), $\rho(m) = 1/2^m$ defines an a.s. discrete measure that coincides with the Dirichlet process (Ferguson, 1973), and $\rho(m) = 1$ defines a continuous singular measure. Moreover, if ρ is such that $\sum_{m=1}^{\infty} \rho(m)^{-1} < \infty$ it guarantees that F is absolutely continuous (Kraft, 1964), e.g., $\rho(m) = m^2, m^3, 2^m, 4^m$.

In practice we need to stop partitioning the space at a finite level M to define a finite tree process. At the lowest level M , we can spread the probability within each set B_{Mj} according to f_0 . In this case the random probability measure defined will have a density of the form

$$f(x) = \prod_{m=1}^M Y_{m, j_m^{(X)}} 2^M f_0(x), \quad (2)$$

for $X \in \mathbb{R}$, and with $j_m^{(X)}$ identifying the set at level m that contains X . This maintains the condition $E(f) = f_0$. We denote a finite Pólya tree process as $\mathcal{PT}_M(\alpha, \rho, F_0)$. Taking $M \rightarrow \infty$ defines a draw from a Pólya tree.

Let us consider a set of functions $\rho(m)$ of the following types:

$$\rho_1(m) = 1/2^m, \quad \rho_2(m) = 1, \quad \rho_3(m) = m^\delta, \quad \text{and} \quad \rho_4(m) = \delta^m, \quad (3)$$

where $\delta > 1$, to define discrete, singular and two absolutely continuous measures, respectively.

To measure “distance” between probability distributions, we concentrate on the Kullback-Leibler divergence, which for densities f and g is defined as

$$\text{KL}(f||g) = \mathbb{E}_f \left[\log \left\{ \frac{f(x)}{g(x)} \right\} \right] = \int \log \left\{ \frac{f(x)}{g(x)} \right\} f(x) dx. \quad (4)$$

3 Properties

3.1 Pólya Trees

If $F \sim \mathcal{PT}_M(\alpha, \rho, F_0)$ then it is not difficult to show that the KL between the centring distribution F_0 and a random draw F is a random variable that does not depend on F_0 , and is given by:

$$\text{KL}(f_0||f) = - \sum_{m=1}^M \sum_{j=1}^{2^m} (\log Y_{mj}) \frac{1}{2^m} - M \log 2. \quad (5)$$

Since the KL divergence measure is asymmetric, we can reverse the role of f and f_0 . In this case the reverse KL divergence becomes:

$$\text{KL}(f||f_0) = \sum_{m=1}^M \sum_{j=1}^{2^m} (\log Y_{mj}) \prod_{k=1}^m Y_{m-k+1, j_{m-k+1}^{(m,j)}} + M \log 2. \quad (6)$$

We now present some results that characterize the first two moments of these divergences.

Proposition 1 *Let $F \sim \mathcal{PT}_M(\alpha, \rho, F_0)$. Then the Kullback-Leibler divergence between f_0 and f , defined in (5), has mean and variance given by*

$$E\{\text{KL}(f_0||f)\} = \sum_{m=1}^M \{\psi_0(2\alpha\rho(m)) - \psi_0(\alpha\rho(m)) - \log 2\}$$

and

$$\text{Var}\{\text{KL}(f_0||f)\} = \sum_{m=1}^M \frac{1}{2^m} \{\psi_1(\alpha\rho(m)) - 2\psi_1(2\alpha\rho(m))\},$$

where $\psi_0(\cdot)$ and $\psi_1(\cdot)$ denote the digamma and trigamma functions respectively¹.

PROOF. The expected value follows by noting that the geometric mean of a beta random variable is $E(\log Y_{mj}) = \psi_0(2\alpha\rho(m)) - \psi_0(\alpha\rho(m))$. For the variance, we use the fact that the random variables Y_{mj} are independent across m , and for the same m , Y_{mj} and Y_{mk} are independent for $|k - j| > 1$. Noting that $\text{Var}(\log Y_{mj}) = \psi_1(\alpha\rho(m)) - \psi_1(2\alpha\rho(m))$ and since $Y_{m,2j} = 1 - Y_{m,2j-1}$, for $j = 1, \dots, 2^{m-1}$, with $\text{Cov}\{\log Y_{m,2j-1}, \log(1 - Y_{m,2j})\} = -\psi_1(2\alpha\rho(m))$, the result follows. \diamond

We now concentrate on the limiting behaviour of the expected KL value as a function of the finite tree level M . For some cases of the function $\rho(\cdot)$ this limit is finite. This is given in the following corollary.

Corollary 1 *Let $\mathcal{E}_M := E\{\text{KL}(f_0||f)\}$, given in Proposition 1, to make explicit the dependence on the maximum level M . For the families $\rho_3(m)$ and $\rho_4(m)$ in expression (3), the limit of the expected KL divergence, as $M \rightarrow \infty$, is bounded respectively by:*

$$\lim_{M \rightarrow \infty} \mathcal{E}_M \leq \frac{1}{4\alpha} \zeta(\delta) + \frac{1}{\alpha^2} \zeta(\delta^2) \quad (7)$$

$$\lim_{M \rightarrow \infty} \mathcal{E}_M \leq \frac{\alpha(\delta + 1) + 4}{4\alpha^2(\delta^2 - 1)} \quad (8)$$

where δ is defined as in (3), and $\zeta(\delta) = \sum_{n=1}^{\infty} n^{-\delta}$, is the Riemann zeta function.

PROOF. The digamma function can be expanded as: $\psi_0(x) = \log x - (1/2)x^{-1} - \mathcal{O}(x^{-2})$, from which these inequalities follow. \diamond

Taking instead the reverse KL, we have the following properties.

¹The digamma function is defined as the logarithmic derivative of the gamma function, i.e. $\psi_0(x) = \frac{d}{dx} \log \Gamma(x) = \frac{\Gamma'(x)}{\Gamma(x)}$. In similar fashion, the trigamma function is defined as the second derivative.

Proposition 2 *Let $F \sim \mathcal{PT}_M(\alpha, \rho, F_0)$. Then the Kullback-Leibler divergence between f and f_0 , defined in (6), has mean and variance given by*

$$E\{\text{KL}(f||f_0)\} = \sum_{m=1}^M \{\psi_0(\alpha\rho(m) + 1) - \psi_0(2\alpha\rho(m) + 1) + \log 2\}$$

and

$$\text{Var}\{\text{KL}(f||f_0)\} = A + B,$$

where

$$\begin{aligned} A &= \sum_{m=1}^M \left[\left\{ \prod_{k=1}^m \left(\frac{\alpha\rho(k) + 1}{2\alpha\rho(k) + 1} \right) \right\} \lambda_5(m) - \left(\frac{1}{2} \right)^m \lambda_2^2(m) \right], \\ B &= \sum_{m=1}^M \left(\left(\frac{\alpha\rho(m)}{2\alpha\rho(m) + 1} \right) \left\{ \prod_{k=1}^{m-1} \left(\frac{\alpha\rho(k) + 1}{2\alpha\rho(k) + 1} \right) \right\} \lambda_6(m) - \left(\frac{1}{2} \right)^m \lambda_2^2(m) \right. \\ &\quad + \sum_{j=1}^{m-1} \left[\left(\frac{\alpha\rho(j)}{2\alpha\rho(j) + 1} \right) \left\{ \prod_{k=1}^{j-1} \left(\frac{\alpha\rho(k) + 1}{2\alpha\rho(k) + 1} \right) \right\} \lambda_2^2(m) - \left(\frac{1}{2} \right)^j \lambda_2^2(m) \right] \\ &\quad + 2 \left\{ \prod_{k=1}^{m-1} \left(\frac{\alpha\rho(k) + 1}{2\alpha\rho(k) + 1} \right) \right\} \sum_{j=m+1}^M \left\{ \left(\frac{\alpha\rho(m) + 1}{2\alpha\rho(m) + 1} \right) \lambda_3(m) \lambda_2(j) \right. \\ &\quad \left. \left. + \left(\frac{\alpha\rho(m)}{2\alpha\rho(m) + 1} \right) \lambda_4(m) \lambda_2(j) - \lambda_2(m) \lambda_2(j) \right\} \right) \end{aligned}$$

with

$$\begin{aligned} \lambda_2(m) &= \psi_0(\alpha\rho(m) + 1) - \psi_0(2\alpha\rho(m) + 1), \\ \lambda_3(m) &= \psi_0(\alpha\rho(m) + 2) - \psi_0(2\alpha\rho(m) + 2), \\ \lambda_4(m) &= \psi_0(\alpha\rho(m) + 1) - \psi_0(2\alpha\rho(m) + 2), \\ \lambda_5(m) &= \psi_1(\alpha\rho(m) + 2) - \psi_1(2\alpha\rho(m) + 2) + \{\psi_0(\alpha\rho(m) + 2) - \psi_0(2\alpha\rho(m) + 2)\}^2, \\ \lambda_6(m) &= \{\psi_0(\alpha\rho(m) + 1) - \psi_0(2\alpha\rho(m) + 2)\}^2 - \psi_1(2\alpha\rho(m) + 2). \end{aligned}$$

PROOF. The expected value follows by using independence properties and by noting that $E\{(\log Y_{mj})Y_{mj}\} = \lambda_2(m)/2$. For the variance, we first bring

the variance operator within the sum by splitting it into the sum of variances of each element plus the sum of covariances². \diamond

Figures 1 and 2 respectively illustrate the behaviour of the mean and standard deviation, as a function of the truncation level M for the two KL measures (5) (empty dots) and (6) (solid dots). The four panels in each figure correspond to choices of $\rho(m) = 1/2^m, 1, m^\delta, \delta^m$, as given in (3). In all cases we use $\alpha = 1$, and $\delta = 2$ (the so-called canonical choice). The plots show that $E\{KL(f_0||f)\} \geq E\{KL(f||f_0)\}$ for all M and for all functions ρ . Apart from the singular continuous case, $\rho_2(m) = 1$, the variances of $KL(f_0||f)$ are also larger than those of $KL(f||f_0)$.

We see that for the case of $\rho_1(m) = 1/2^m$, which corresponds to the Dirichlet process, the mean value of the KL and the reverse KL diverge to infinity as $M \rightarrow \infty$ ³. The KL (5) increases at an exponential rate whereas for the reverse KL (6) the growth rate is constant. As for the standard deviations, that of the KL also diverges as $M \rightarrow \infty$, however, that of the reverse KL converges.

The precision function $\rho_2(m) = 1$, which defines a singular continuous random distribution (Ferguson, 1974), has asymptotic constant expected values for both KL and reverse KL in the limit of M . The variance of the KL converges to a finite value when $M \rightarrow \infty$, but for the reverse KL the variance increases at a constant rate as a function of M . In the case of the two continuous processes, obtained with precision functions ρ_3 and ρ_4 , the expected values for KL and the reverse KL converge in the limit, as given by the upper bounds in Corollary 1. Interestingly, the variances for the two KL divergences are asymptotically constant.

These results give a precise interpretation of any choice of parametrisation of a Pólya tree, summarised in the choice to the two parameters (α, δ) ⁴. The conventional method for informing the parametrisation of a Pólya tree

²The variance of each element is defined in terms of first and second moments and rely on independence properties to compute them. Working out the algebra with patience and noting that $E\{(\log Y_{mj})Y_{mj}\} = \lambda_2(m)/2$, $E\{(\log Y_{mj})Y_{mj}^2\} = \frac{1}{2} \left(\frac{\alpha\rho(m)+1}{2\alpha\rho(m)+1} \right) \lambda_3(m)$, $E\{(\log Y_{mj})Y_{mj}(1 - Y_{mj})\} = \frac{1}{2} \left(\frac{\alpha\rho(m)}{2\alpha\rho(m)+1} \right) \lambda_4(m)$, $E\{(\log Y_{mj})^2 Y_{mj}^2\} = \frac{1}{2} \left(\frac{\alpha\rho(m)+1}{2\alpha\rho(m)+1} \right) \lambda_5(m)$, and $E\{(\log Y_{mj}) \log(1 - Y_{mj}) Y_{mj}(1 - Y_{mj})\} = \frac{1}{2} \left(\frac{\alpha\rho(m)}{2\alpha\rho(m)+1} \right) \lambda_6(m)$, the result is obtained.

³Figure 1 appears to show that $E\{KL(f||f_0)\}$ remains constant, but this is an artefact due to the scale.

⁴Here we only consider the class of functions $\rho_3(m) = m^\delta$.

process is that the choice of δ is unimportant, with default value 2, and thus the choice of α completely controls the variety of draws from the process. However, these two parameters are confounded and should not be chosen independently. As shown in figure 4, the expected KL is dependent on both parameters, although choices of α near to zero mean that the exponent δ has little effect on the expected KL and its variance.

3.2 Frequentist and Bayesian “bootstrap”

Let us now consider the setting where f_0 is a discrete density with n atoms $\{\xi_1, \dots, \xi_n\}$, i.e., $f_0(x) = \sum_{i=1}^n p_i \delta_{\xi_i}(x)$, with $p_i > 0$ for all $i = 1, \dots, n$ and $\sum_{i=1}^n p_i = 1$. Let $\mathbf{w} = (w_1, \dots, w_n)$ be random weights such that $w_i \geq 0$ and $\sum_{i=1}^n w_i = 1$ almost surely. Let f be a random distribution defined as a reweighing of the atoms of f_0 with the random weights \mathbf{w} . In notation, $f(x) = \sum_{i=1}^n w_i \delta_{\xi_i}(x)$.

The Kullback-Leibler divergence between f_0 and f does not depend on the atoms locations and is given by:

$$\text{KL}(f_0||f) = \sum_{i=1}^n p_i \log \left(\frac{p_i}{w_i} \right), \quad (9)$$

and the reverse Kullback-Leibler has the form

$$\text{KL}(f||f_0) = \sum_{i=1}^n w_i \log \left(\frac{w_i}{p_i} \right). \quad (10)$$

If we take f_0 to have uniform weights, such as when X is a random sample from some population and F_0 represents the empirical CDF, we first highlight an important property in the relationship between the divergences (9) and (10).

Proposition 3 *Consider the KL divergences (9) and (10). If $p_i = 1/n$ for $i = 1, \dots, n$, then for any given re-weighing vector \mathbf{w} taken from the simplex $\mathcal{Q}_n := \{w : w_i \geq 0, \sum_{i=1}^n w_i = 1\}$ we have that*

$$\text{KL}(f_0||f) \geq \text{KL}(f||f_0).$$

PROOF. Let $h(\mathbf{w}) := \text{KL}(f_0||f) - \text{KL}(f||f_0)$. Using expressions (9) and (10), $h(w)$ becomes $h(\mathbf{w}) = -\sum (1/n + w_i) \log(w_i)$. We note that $h(\mathbf{w}) = 0$ at

$\mathbf{w}^* = (1/n, \dots, 1/n)$ and is infinite on all the simplex boundaries. Moreover, h is convex and by straightforward differentiation we see that $h''(\mathbf{w}^*)$ is positive. The result follows. \diamond

This result is consistent with the results from previous section. However, in this particular discrete setting $\text{KL}(f_0||f)$ dominates $\text{KL}(f||f_0)$.

Taking for instance $n\mathbf{w} \sim \text{Mult}(n, \mathbf{p})$ ⁵, a multinomial distribution with n trials and n categories with probability of success $\mathbf{p} = (p_1, \dots, p_n)$, means that the random f 's will be centred at f_0 . It is not difficult to show that $E(f) = f_0$. Note that if $p_i = 1/n$ for $i = 1, \dots, n$ this choice of distribution for the weights \mathbf{w} coincides with the frequentist bootstrap (Efron, 1979) for which the atoms $\{\xi_i\}$ are replaced by i.i.d. random variables $\{X_i\}$.

We note that the KL divergence (9) will not in general be defined, as w_i can be zero. In fact, for large n and for $p_i = 1/n$ in the previous multinomial choice, approximately one third of the weights will be zero. However, $0 \log 0$ is defined by convention as 0, so the reverse KL (10) is well defined.

Proposition 4 *The expected value of the Kullback-Leibler between a “bootstrap” draw f , with $n\mathbf{w} \sim \text{Mult}(n, \mathbf{p})$, and its centring distribution f_0 , defined in (10), has the following upper bound:*

$$E\{\text{KL}(f||f_0)\} \leq \sum_{i=1}^n p_i \log \left(p_i + \frac{1-p_i}{n} \right) - H(\mathbf{p}) \quad (11)$$

where $H(\mathbf{p}) = \sum_{i=1}^n p_i \log p_i$, the entropy of the vector \mathbf{p} . For the special case when $p_i = 1/n$, we have $E\{\text{KL}(f||f_0)\} \leq \log(2 - 1/n) \leq \log 2$

PROOF.

$$E\{\text{KL}(f||f_0)\} = \sum_{i=1}^n E\{w_i \log w_i\} - \sum_{i=1}^n E\{w_i\} \log p_i.$$

Working on the individual expected values,

$$E(w_i \log w_i) = \sum_{k=1}^n \binom{n}{k} p_i^k (1-p_i)^{n-k} \left(\frac{k}{n} \right) \log \left(\frac{k}{n} \right).$$

⁵We use this notation to emphasise the fact that \mathbf{w} represents a random probability mass function, but taking values on the set $\{0, 1/n, 2/n, \dots, 1\}$. A factor of n is needed for the vector to be distributed according to a multinomial distribution.

From which we get $E(w_i \log w_i) = (1/n)E\{\log((v_i + 1)/n)\}$, with $v_i \sim \text{Bin}(n-1, 1/n)$. Using Jensen's inequality we get $E\{w_i \log w_i\} \leq p_i \log(p_i + (1 - p_i)/n)$. Substituting this into the original sum and using $E\{w_i\} = p_i$ gives the result.

◇

An alternative way of making the random f 's to be centred around f_0 is by sampling weights \mathbf{w} from a Dirichlet distribution with parameter vector $\boldsymbol{\beta} = (\beta_1, \dots, \beta_n)$ such that $\beta_i = \alpha_n p_i$, $i = 1, \dots, n$, with $\alpha_n > 0$ a parameter changing as a function of the number of atoms. This is denoted $\mathbf{w} \sim \text{Dir}(\alpha_n \mathbf{p})$. It is straightforward to prove that $E(f) = f_0$, and that the form of α_n parametrises the precision, analogous to the Pólya tree case. If we take $\alpha_n = n$, $p_i = 1/n$ and replace the atoms $\{\xi_i\}$ by i.i.d. random variables $\{X_i\}$, we obtain the original Bayesian bootstrap proposed by Rubin (1981)⁶. Sampling from a Dirichlet with parameter vector $\alpha_n \mathbf{p}$ gives a generalised version of this bootstrap procedure. Ishwaran & Zarepour (2002) considered this model albeit in a different context. In this new setting, both the $\text{KL}(f_0||f)$ and the reverse $\text{KL}(f||f_0)$, given in (9) and (10) respectively, are well defined since $w_i \neq 0$ almost surely. Their expected values and variances can be obtained in closed form as functions of α_n and \mathbf{p} .

Proposition 5 *Let f be a “generalised Bayesian bootstrap” draw around f_0 with weights $\mathbf{w} \sim \text{Dir}(\alpha_n \mathbf{p})$. Then the Kullback-Leibler divergence given in (9) has mean and variance:*

$$E\{\text{KL}(f_0||f)\} = H(\mathbf{p}) - \sum_{i=1}^n p_i \{\psi_0(\alpha_n p_i) - \psi_0(\alpha_n)\}$$

$$\text{Var}\{\text{KL}(f_0||f)\} = \sum_{i=1}^n p_i^2 \psi_1(\alpha_n p_i) - \psi_1(\alpha_n)$$

where ψ_0 and ψ_1 are the digamma and trigamma functions.

PROOF. This result follows from $E(\log w_i) = \psi_0(\alpha_n p_i) - \psi_0(\alpha_n)$ and linearity of expectation. The variance follows from $\text{Var}(\log w_i) = \psi_1(\alpha_n p_i) - \psi_1(\alpha_n)$, and $\text{Cov}(\log w_i, \log w_j) = \psi_1(\alpha_n p_i) \delta_{ij} - \psi_1(\alpha_n)$, where δ_{ij} is the Kronecker delta function taking value 1 when $i = j$ and 0 otherwise. ◇

The limiting behaviour of this expected KL and its variance, as n tends to infinity, can more easily be studied for the special case of $p_i = 1/n$,

⁶It is interesting to note that in the original work they only consider this special case.

$i = 1, \dots, n$. When $\alpha_n = \alpha$, i.e. constant, they both diverge to infinity. In the limit, this is a well known construction of a Dirichlet process, when the atoms are sampled i.i.d. from a baseline measure G . However, if we make α_n grow linearly with n , say $\alpha_n = \alpha n$, then $\lim_{n \rightarrow \infty} E\{\text{KL}(f_0||f)\} = \log(\alpha) - \psi_0(\alpha)$ and $\lim_{n \rightarrow \infty} \text{Var}\{\text{KL}(f_0||f)\} = 0$. These values are obtained by noting that $\psi_0(n)$ behaves like $\log(n)$ for large n . Finally, if we increase the rate at which α_n grows with n , say $\alpha_n = \alpha n^2$, both mean and variance of the KL converge to zero as $n \rightarrow \infty$.

Proposition 6 *Let f be a “generalised Bayesian bootstrap” draw around f_0 with weights $\mathbf{w} \sim \text{Dir}(\alpha_n \mathbf{p})$. Then the Kullback-Leibler divergence given in (10) has mean:*

$$E\{\text{KL}(f||f_0)\} = \sum_{i=1}^n p_i \{\psi_0(\alpha_n p_i + 1) - \psi_0(\alpha_n + 1)\} - H(\mathbf{p}) \quad (12)$$

where $H(\mathbf{p}) := \sum_{i=1}^n p_i \log p_i$ the entropy of the vector \mathbf{p} , and the variance given by

$$\begin{aligned} \text{Var}(\text{KL}(f||f_0)) &= \sum_{i=1}^n \{ \text{Var}(w_i \log w_i) + (\log p_i)^2 \text{Var}(w_i) - 2(\log p_i) \text{Cov}(w_i \log w_i, w_i) \} \\ &+ 2 \sum_{i < j} \{ \text{Cov}(w_i \log w_i, w_j \log w_j) + (\log p_i)(\log p_j) \text{Cov}(w_i, w_j) - 2(\log p_j) \text{Cov}(w_i \log w_i, w_j) \} \end{aligned} \quad (13)$$

where each of the elements are given in the footnote⁷.

PROOF. Note that each $w_i \sim \text{Be}\{\alpha_n p_i, \alpha_n(1 - p_i)\}$ and thus we have that $E(w_i \log w_i) = p_i \{\psi_0(\alpha_n p_i + 1) - \psi_0(\alpha_n + 1)\}$. Using linearity of expectation and substituting this expression we obtain the mean. Using properties of the variance and covariance of sums we get the second part of the result. \diamond

⁷ $\text{Var}(w_i) = p_i(1 - p_i)/(\alpha_n + 1)$, $\text{Cov}(w_i, w_j) = -p_i p_j / (\alpha_n + 1)$, $\text{Var}(w_i \log w_i) = p_i(\alpha_n p_i + 1)/(\alpha_n + 1) \{\psi_1(\alpha_n p_i + 2) - \psi_1(\alpha_n + 2) + [\psi_0(\alpha_n p_i + 2) - \psi_0(\alpha_n + 2)]^2\} - p_i^2 \{\psi_0(\alpha_n p_i + 1) - \psi_0(\alpha_n + 1)\}^2$, $\text{Cov}(w_i \log w_i, w_i) = p_i(\alpha_n p_i + 1)/(\alpha_n + 1) \{\psi_0(\alpha_n p_i + 2) - \psi_0(\alpha_n + 2)\} - p_i^2 \{\psi_0(\alpha_n p_i + 1) - \psi_0(\alpha_n + 1)\}$, $\text{Cov}(w_i \log w_i, w_j) = p_i p_j \{-\psi_0(\alpha_n p_i + 1)/(\alpha_n + 1) + \psi_0(\alpha_n + 1) - \alpha_n \psi_0(\alpha_n + 2)/(\alpha_n + 1)\}$, $\text{Cov}(w_i \log w_i, w_j \log w_j) = \alpha_n p_i p_j / (\alpha_n + 1) [\{\psi_0(\alpha_n p_i + 1) - \psi_0(\alpha_n + 2)\} \{\psi_0(\alpha_n p_j + 1) - \psi_0(\alpha_n + 2)\} - \psi_1(\alpha_n + 2)] - p_i p_j \{\psi_0(\alpha_n p_i + 1) - \psi_0(\alpha_n + 1)\} \{\psi_0(\alpha_n p_j + 1) - \psi_0(\alpha_n + 1)\}$.

Similarly to the previous case, if we take $p_i = 1/n$ and $\alpha_n = \alpha n$, when $n \rightarrow \infty$ then $E\{\text{KL}(f||f_0)\} \rightarrow \psi_0(\alpha + 1) - \log(\alpha)$. It is possible to show analytically that each term in (13) goes to zero as $n \rightarrow \infty$, but this can also be seen using the relation between the two KLs given in Proposition 3, and noting that the variance involves a monotonic transformation, hence we have that $\text{Var}\{\text{KL}(f_0||f)\} \geq \text{Var}\{\text{KL}(f||f_0)\}$. From the previous result it follows that $\lim_{n \rightarrow \infty} \text{Var}\{\text{KL}(f||f_0)\} = 0$ for these choices of p_i and α_n .

In Figure 5 we compare the expected value and variance of both KL and reverse KL for $p_i = 1/n$ and different values of α_n as a function of n . The first column corresponds to $\alpha_n = 1$, the second column to $\alpha_n = n$ and the third to $\alpha_n = n^2$, which induce high, moderate and small variance in the \mathbf{w} respectively. In accordance to what we have proved, the expected value and variance of $\text{KL}(f_0||f)$ are larger than those of $\text{KL}(f||f_0)$, and their limiting behaviours can also be assessed from the graphs.

If we replace the atoms $\{\xi_i\}$ by i.i.d. random variables $\{X_i\}$ from a distribution G and take $p_i = 1/n$ for $i = 1, \dots, n$, then f_0 represents the empirical density for the random variables X_i 's and f represents a random process centred around the empirical. Ishwaran & Zarepour (2002) considered exactly this random probability process and derived results for the limiting behaviour for a variety of choices of α_n (see Theorem 3, page 948). Let F be the cdf associated to f . When $\alpha_n = \alpha$, then F is distributed according to a Dirichlet process $\mathcal{DP}(\alpha, G)$, in the limit as $n \rightarrow \infty$. If $\alpha_n = \alpha n$, then we have almost sure weak convergence of F to G , as $n \rightarrow \infty$. For the third case considered here, $\alpha_n = \alpha n^2$, F converges in probability to G , as $n \rightarrow \infty$.

The case where $\alpha_n = \alpha n$ is of particular interest. Although we have weak convergence of $F \rightarrow G$, the random distribution does not converge in KL divergence. In other words, although functionals of f tend to the functionals of f_0 , the KL divergence between the two densities remains non zero. This becomes apparent when considering the random quantity nw_i , which comes into the equation (9), whose variance becomes asymptotically $1/\alpha$, as $n \rightarrow \infty$. Convergence in Kullback-Leibler is a strong statement, stronger than convergence of functionals and L_1 convergence. A more intuitive illustration is the posterior convergence of two Dirichlet processes with different baseline measures (that have the same support). By posterior consistency, both will weakly converge to the same measure, but their L_1 divergence will remain finite and their KL divergence will remain infinite.

4 Discussion

This note explores properties of the KL and reverse KL of draws F from some classical random probability models with respect to their centring distribution F_0 . These properties become relevant when applying a particular process as a modelling tool. For example, draws from the Dirichlet process prior have divergent expected KL (obtained in our Pólya tree setting with ρ_1 in (3) and $M \rightarrow \infty$, and also obtained in the Bayesian bootstrap setting with $\alpha_n = \alpha$ and $n \rightarrow \infty$). Therefore we can say that any random draw taken from a Dirichlet process prior is completely “different” from the baseline distribution as measured in terms of the KL divergence, regardless the value of α . This is also a surprising result but in accordance with the full support property of the Dirichlet process⁸.

Our key result concerns the Pólya tree prior. In the majority of applications, it is usually constructed in its continuous version, i.e. the precision function ρ satisfies the continuity property, for example ρ_3 and ρ_4 as given in (3). In these cases, the first two moments of the distance (in KL units) of the draws from their centring measure is given as an explicit function of the truncation level M and the precision function $\alpha\rho(m)$. Therefore the specification of the truncation level M , precision parameter α , and precision function ρ are all highly important, with careless choices leading to a prior overly concentrated around f_0 . The vast majority of applications with Pólya tree priors use the family $\rho_3(m)$ with choice of exponent $\delta = 2$. In Figure 3 we show that using $\rho_3(m)$ with a choice of $\delta = 1.01$ (empty dots) gives greater gains in expected KL as M is increased as compared to those obtained for the standard choice of $\delta = 2$ and decreasing the parameter α . The concentration around the baseline measure is highly sensitive to this choice of exponent, thus questioning the “sensible canonical choice” of $\delta = 2$ given by Lavine (1992).

Moreover, in practice Pólya trees are used in their finite versions, that is, finite M . In such cases the choice of M has been done with a rule of thumb (e.g. Hanson, 2006), say $M = \log_2(n)$ with n being the data sample size. The authors note ‘a law of diminishing returns’ when increasing the truncation level from $M \rightarrow M + 1$. Our study confirms this by plotting the diversity of draws as measured in KL against M , and these findings suggest that a

⁸Stated in Ferguson (1973), saying that any fixed density (measure) g absolutely continuous with respect to f_0 can be arbitrarily approximated with a draw f from a Dirichlet process.

Pólya tree prior with as a low as $M = 4$ and $\rho(m) = 2^m$ can produce random draws that are equally far from the centring distribution as with a larger M (see two bottom panels in Figures 1 and 2). If it desired to make proper use of finite nature of the tree, the various possibilities in specification of the precision function ρ within families that satisfy the continuity property should be used.

In the discrete setting, we can always see f_0 as the empirical density obtained from a sample of size n taken from a continuous density. This is often the case when characterising a posterior distribution in Bayesian analysis, for example via MCMC sampling (e.g. Gelman *et al.*, 2013). One lesson from this work, is that by increasing n , the variance of the reverse KL in the frequentist bootstrap, and the variance of the KL and reverse KL for the Bayesian bootstrap with $\alpha_n = \alpha n$, converge to zero. This implies that for large n a frequentist or Bayesian bootstrap draw lies below $\log(2)$ and exactly at $\log(\alpha) - \psi_0(\alpha)$ or $\psi_0(\alpha + 1) - \log(\alpha)$ in KL units, respectively.

Acknowledgements

We are grateful to Judith Rousseau for helpful comments. Watson is supported by the Industrial Doctoral Training Centre (SABS-IDC) at Oxford University and Hoffman-La Roche. This work was done whilst Nieto-Barajas was visiting the Department of Statistics at the University of Oxford. He is supported by *Asociación Mexicana de Cultura, A.C.*–Mexico. Holmes gratefully acknowledges support for this research from the EPSRC and the Medical Research Council.

References

- Bernardo, J.M., & Smith, A.F.M. 1994. *Bayesian Theory*. Wiley Series in Probability and Mathematical Statistics. John Wiley & Sons.
- Cover, T.M., & Thomas, J.A. 1991. *Elements of information theory*. Wiley, New York.
- Efron, B. 1979. Bootstrap methods: another look at the jackknife. *The Annals of Statistics*, 1–26.

- Ferguson, T.S. 1973. A Bayesian analysis of some nonparametric problems. *The Annals of Statistics*, 209–230.
- Ferguson, T.S. 1974. Prior distributions on spaces of probability measures. *The Annals of Statistics*, 615–629.
- Gelman, A., Carlin, J.B., Stern, H.S., Dunson, D.B., Vehtari, A., & Rubin, D.B. 2013. *Bayesian data analysis*. Chapman and Hall, Boca Raton.
- Ghosh, J.K., & Ramamoorthi, R.V. 2003. *Bayesian nonparametrics*. Vol. 1. Springer.
- Hanson, T., & Johnson, W.O. 2002. Modeling regression error with a mixture of Pólya trees. *Journal of the American Statistical Association*, **97**(460).
- Hanson, T.E. 2006. Inference for mixtures of finite Pólya tree models. *Journal of the American Statistical Association*, **101**(476).
- Hjort, N.L., Holmes, C., Müller, P., & Walker, S.G. 2010. *Bayesian Nonparametrics*. Cambridge University Press.
- Ishwaran, H., & Zarepour, M. 2002. Dirichlet prior sieves in finite normal mixtures. *Statistica Sinica*, **12**(3), 941–963.
- Karabatsos, G. 2006. Bayesian nonparametric model selection and model testing. *Journal of Mathematical Psychology*, **50**(2), 123–148.
- Kraft, C.H. 1964. A class of distribution function processes which have derivatives. *Journal of Applied Probability*, **1**(2), 385–388.
- Kullback, S. 1997. *Information theory and statistics*. Courier Dover Publications.
- Kullback, S., & Leibler, R.A. 1951. On information and sufficiency. *The Annals of Mathematical Statistics*, 79–86.
- Lavine, M. 1992. Some aspects of Pólya tree distributions for statistical modelling. *The Annals of Statistics*, 1222–1235.
- Muliere, P., & Walker, S. 1997. A Bayesian Non-parametric Approach to Survival Analysis Using Pólya Trees. *Scandinavian Journal of Statistics*, **24**(3), 331–340.

- Müller, P., & Quintana, F.A. 2004. Nonparametric Bayesian data analysis. *Statistical science*, 95–110.
- Nieto-Barajas, L. E, & Müller, P. 2012. Rubbery Pólya Tree. *Scandinavian Journal of Statistics*, **39**(1), 166–184.
- Rubin, D.B. 1981. The Bayesian bootstrap. *The Annals of Statistics*, **9**(1), 130–134.
- Walker, S. G., Damien, P., Laud, P. W., & Smith, A.F.M. 1999. Bayesian Nonparametric Inference for Random Distributions and Related Functions. *Journal of the Royal Statistical Society. Series B (Statistical Methodology)*, **61**(3), pp. 485–527.
- Walker, S.G., & Mallick, B.K. 1997. Hierarchical generalized linear models and frailty models with Bayesian nonparametric mixing. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, **59**(4), 845–860.

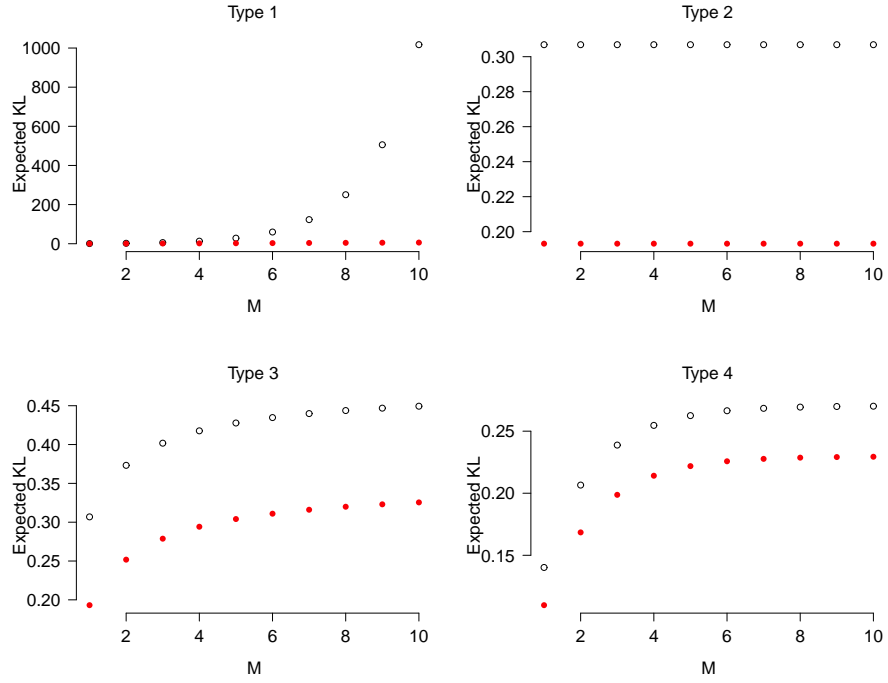


Figure 1: Comparison between expected values of KL for different values of M . $E\{KL(f_0||f)\}$ (empty dots) and $E\{KL(f||f_0)\}$ (solid dots). Type 1 to 4 denote the different ρ functions as in (3).

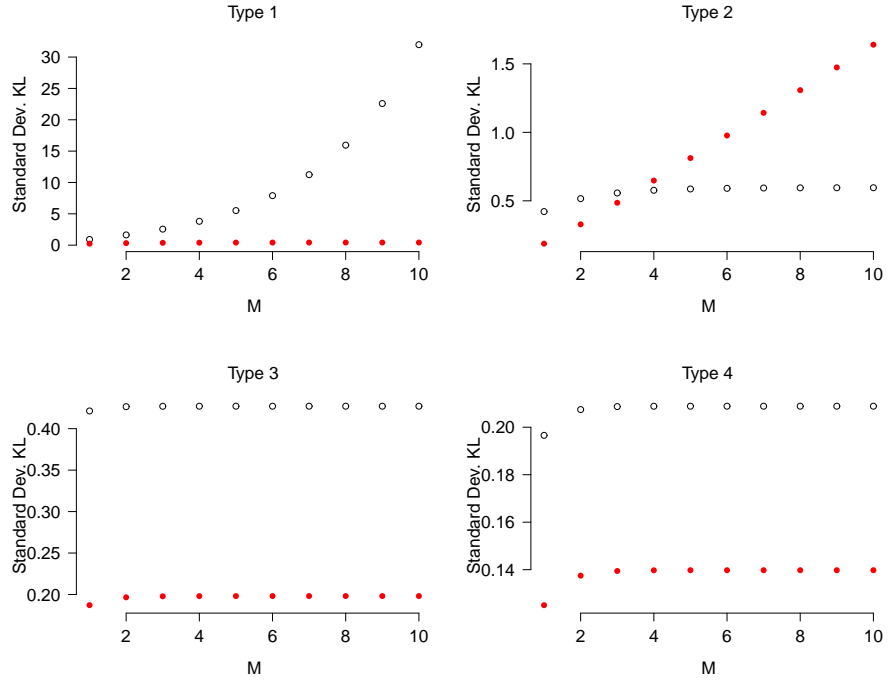


Figure 2: Comparison between standard deviations of KL for different values of M . $\sqrt{\text{Var}\{KL(f_0||f)\}}$ (empty dots) and $\sqrt{\text{Var}\{KL(f||f_0)\}}$ (solid dots). Type 1 to 4 denote the different ρ functions as in (3).

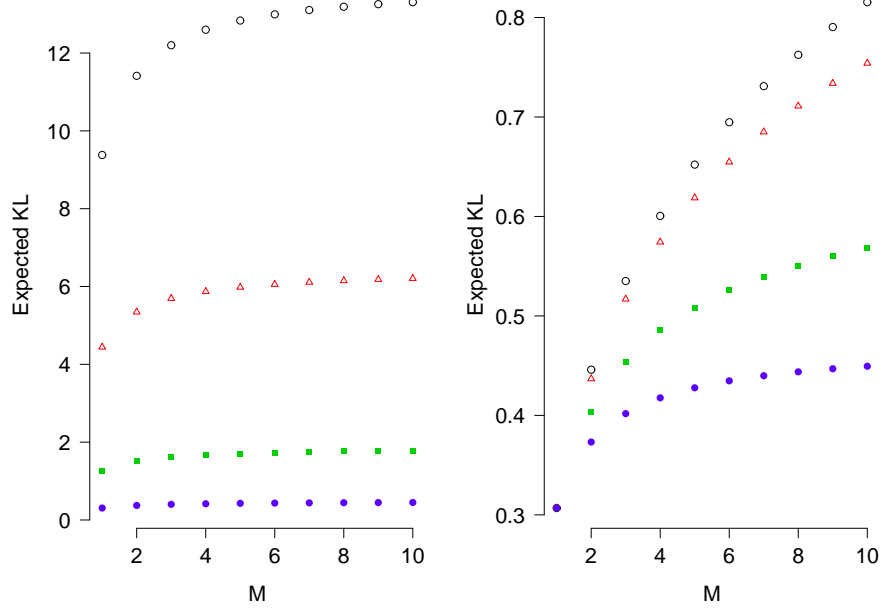


Figure 3: Expected $\text{KL}(f_0||f)$ for varying α (left panel) and $\rho(m) = m^\delta$ (right panel). Left: $\alpha = 0.05$ (empty dots), $\alpha = 0.1$ (triangles), $\alpha = 0.3$ (squares), $\alpha = 1$ (solid dots). Right: $\delta = 1.01$ (empty dots), $\delta = 1.1$ (triangles), $\delta = 1.5$ (squares), $\delta = 2$ (solid dots).

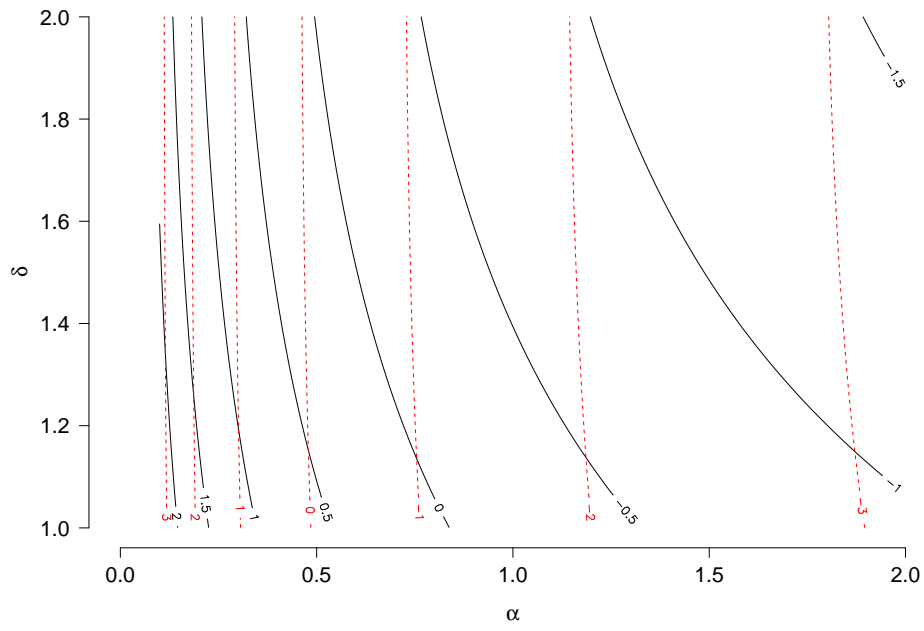


Figure 4: Overlaid contour plots of the log-expected KL (black lines) and the log-variance of the KL (dashed red lines) as functions of the two parameters (α, δ) , at regular intervals of $1/2$ and 1 respectively, of draws from a Pólya tree process with truncation level $M = 10$.

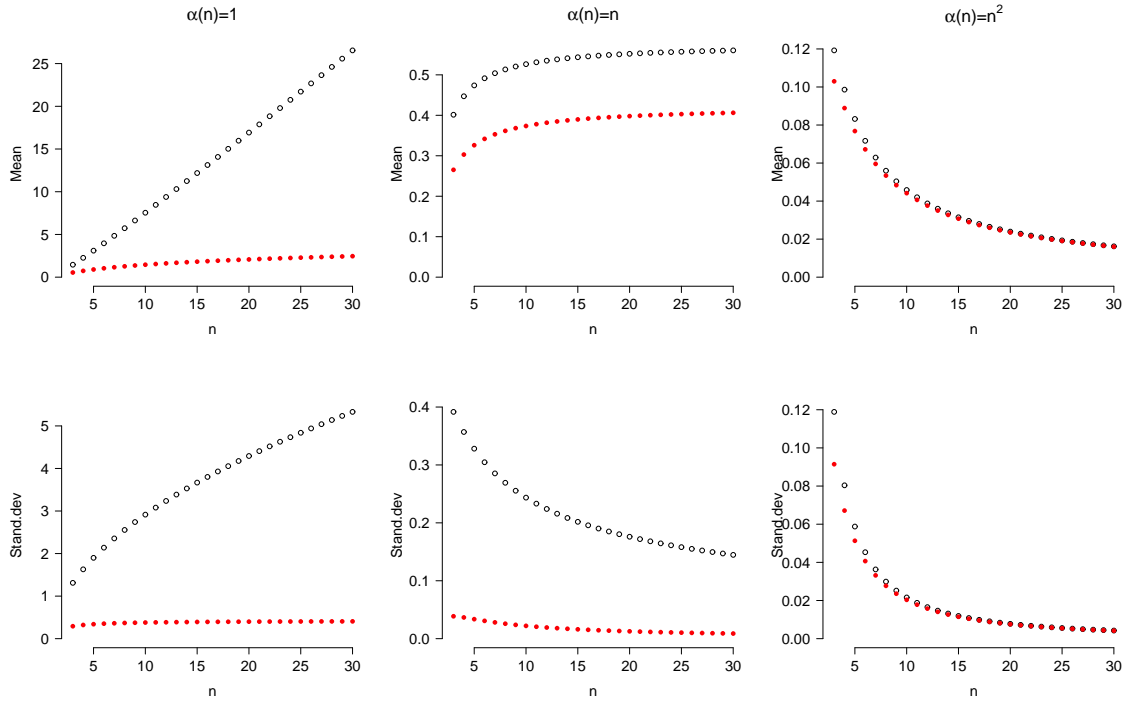


Figure 5: Expected value (top row) and standard deviation (bottom row) of $KL(f_0||f)$ (black empty dots) and $KL(f||f_0)$ (red solid dots). In columns from left to right: $\alpha_n = 1$, $\alpha_n = n$ and $\alpha_n = n^2$.